The Continued Saga of DB-IR Integration

Ricardo Baeza-Yates

Center for Web Research
Dept. of Computer Science
University of Chile
Blanco Encalada 2120
Santiago, Chile
rbaeza@dcc.uchile.cl

Mariano Consens

Mechanical & Industrial Engineering Dept.
University of Toronto
5 King's College Road
Toronto, Ontario
Canada, M5S 3G8
consens@mie.utoronto.ca

Content

The world of data has been developed from two main points of view: the structured relational data model and the unstructured text model. The two distinct cultures of database and information retrieval now have a natural meeting place in the Web with its semi-structured XML model. As web-style searching becomes an ubiquitous tool, the need for integrating these two viewpoints becomes even more important.

The tutorial will provide an overview of the different issues and approaches put forward by the IR and DB communities and survey the DB-IR integration efforts. Both earlier proposals as well as recent ones (in the context of XML in particular) will be discussed. A variety of application scenarios for DB-IR integration will be covered.

The target audience of this tutorial includes researchers in database systems, as well as developers of Web and database/information retrieval applications. Our goal is that at the end there is a clear view of the main problems and pitfalls that should be avoided.

Short Bios

Ricardo Baeza-Yates is professor and chair of the CS department of the University of Chile. He is also director of the Center for Web Research. As professional volunteer, he is member of the IEEE CS Board of Governors for 2002-04; president of CLEI, a Latin American association of CS departments; and coordinator of the Iberoamerican cooperation program in Electronics and Informatics. His research interests include information retrieval, algorithms, and information visualization. He is co-author of the popular textbook *Modern Information Retrieval*, published in 1999 by Addison-Wesley, as well as co-author or co-editor of other books in algorithms and data structures or information retrieval, in addition to more than 50 publications in journals and over 100 other papers. Ricardo received the BSc in CS in 1983, the MSc in CS and the professional title in EE in 1985, and the MEng in EE in 1986 from the University of Chile. He received his PhD in CS from the U. of Waterloo, Canada, in 1989. He has been the president of the Chilean CS Society for 1992-95 and 1997-98. During 1993, he received the Organization of American States award for young researchers in exact sciences. In 1994 he received the award to the best engineering research from the Institute of Engineers of Chile. In 1997 with two Brazilian colleagues obtained the COMPAQ prize to the best Brazilian research article in CS. In 2002 he was appointed to the Chilean Academy of Sciences, being the first person from CS to achieve this position in Chile.

Mariano Consens research interests are in the areas of Data Management Systems and the Web, with a current focus on XML searching, autonomic systems and pervasive computing. He has over 20 publications and two patents, including journal publications selected from best conference papers. Mariano received his PhD and MSc degrees in Computer Science from the University of Toronto. He also holds a Computer Systems Engineer degree from the Universidad de la Republica, Uruguay. Consens has been a faculty member in Information Engineering at the MIE Department, University of Toronto, since 2003. Before that, he was research faculty at the School of Computer Science, University of Waterloo, from 1994 to 1999. In addition, he has been active in the software industry as a founder and Director of Freedom Intelligence (a query engine provider), as the CTO of Classwave Wireless (a Bluetooth software infrastructure company) and as a technology advisor for Xign (an electronic payment systems supplier), OpenText (an early web search engine turned knowledge management software vendor) and others.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment.

Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004